Wilcoxon-type Generalized Bayesian Information Criterion

By LAN WANG

School of Statistics, University of Minnesota,

313 Ford Hall, 224 Church Street S.E., Minneapolis, Minnesota 55455, U.S.A.

lan@stat.umn.edu

Summary

We develop a generalized Bayesian information criterion for regression model selection. The new criterion relaxes the usually strong distributional assumption associated with Schwarz's BIC by adopting a Wilcoxon-type dispersion function and appropriately adjusting the penalty term. We establish that the Wilcoxon-type generalized BIC preserves the consistency property of Schwarz's BIC without the need to assume a parametric likelihood. We also show that it outperforms Schwarz's BIC with heavier-tailed data in the sense that asymptotically it can yield substantially smaller L_2 risk. On the other hand, when the data are normally distributed, both criteria have similar L_2 risk. The new criterion function is convex and can be conveniently computed via existing statistical software. Our proposal provides a flexible yet highly efficient alternative to Schwarz's BIC; at the same time, it broadens the scope of Wilcoxon inference, which has played a fundamental role in classical nonparametric analysis.

Some key words: BIC; Bayesian information criterion; Consistency of model selection; Heavier-tailed distribution; L_2 risk; Rank; Wilcoxon inference.

1. Introduction

There has been much effort to develop variants of Schwartz's (1978) BIC in order to handle increasingly complex data structures. For example, Volinsky & Raftery (2000)

adapted BIC to censored survival models; Broman & Speed (2002) developed BIC for the problem of identification of quantitative loci in experimental crosses; Konishi et al. (2004) extended BIC to choose the smoothing parameter and the number of basis functions in radial basis function networks; Siegmund (2004) modified BIC for changepoint-like problems with an interesting application in genetic linkage analysis; and M. J. Bayarri and her coauthors in their recent on-going work has suggested interesting extensions of BIC in several directions, including the case where the model complexity grows with the sample size.

Given a class of candidate models, BIC favours the model that minimizes

$$-2 \log L(\hat{\theta}|D) + p \log n, \tag{1}$$

where $L(\hat{\theta}|D)$ is the maximized likelihood function of the given model, D represents the data, p is the number of free parameters and n denotes the sample size. In the context of linear regression, which is the focus of this paper, BIC often takes the following familiar form

$$\text{RSS} + \hat{\sigma}^2 p \log n, \tag{2}$$

where RSS is the residual sum of squares from the least squares fit, and $\hat{\sigma}^2$ is an estimator of the error variance, which is usually computed from the full model. When computing BIC for regression-model selection, most statistical software packages adopt the form (2) because it can be obtained directly from standard least-squares regression output. However, the derivation of (2) from (1) relies on the assumption of normality. With normal random errors, Nishii (1984) established that, under mild regularity conditions, (2) leads to the most parsimonious correct model with probability approaching one. This is known as the consistency property of BIC.

Schwartz's BIC requires an unambiguous specification of a parametric distribution. This sometimes seriously limits its application. Moreover, when the underlying probability distribution is misspecified, the above mentioned consistency property is likely to break down.

To overcome this drawback, this paper proposes a new regression model selection procedure called the Wilcoxon-type generalized Bayesian information criterion.

2. Wilcoxon-type rank regression

Before we define the Wilcoxon-type generalized BIC, we first briefly introduce Wilcoxon rank regression.

Consider a multiple linear regression model $Y = \alpha 1_n + X\beta + \epsilon$, where Y is an $n \times 1$ vector of responses, α is the intercept, 1_n is an $n \times 1$ vector of ones, X is an $n \times p$ matrix of covariates which without loss of generality is assumed to be centred, β is a $p \times 1$ vector of unknown parameters, and ϵ is an $n \times 1$ vector of independent, identically distributed random errors with probability density function $f(\cdot)$. The Wilcoxon rank estimator of β minimizes

$$W_n(\beta) = \sqrt{12} \sum_{i=1}^{n} \left\{ \frac{R(Y_i - X_i'\beta)}{n+1} - \frac{1}{2} \right\} (Y_i - X_i'\beta), \tag{3}$$

where X_i' is the ith row of X, and $R(Y_i - X_i'\beta)$ denotes the rank of $Y_i - X_i'\beta$ among $Y_1 - X_1'\beta, \ldots, Y_n - X_n'\beta$. This estimator, which was proposed by Jaeckel (1972), is asymptotically equivalent to the rank estimator of Jureăková (1971).

The objective function (3) is a nonnegative convex function of β and provides a robust measure of the dispersion of the residuals. McKean & Schrader (1980) and Hettmansperger & McKean (1983) further revealed an intuitive geometric interpretation of (3); the above minimization is analogous to the least squares procedure except that the Euclidean norm is substituted by a Wilcoxon-type rank norm. Under the assumptions listed in the Appendix, the Wilcoxon rank estimator is asymptotically normal, robust and highly efficient. For a comprehensive presentation of rank-based analysis of linear models, we refer to Hettmansperger & McKean (1998).

3. Wilcoxon-type generalized BIC

3·1. *Definition*

The problem of variable selection is to identify a subset of the covariates that can describe the information in the data adequately. In what follows, we formally define the Wilcoxon-type generalized BIC.

We begin by indexing each candidate model by a p-dimension binary vector $\nu = (\nu_1, \ldots, \nu_p)'$, where ν_i is one if x_i belongs to the candidate model and is zero otherwise. The total number of ones in ν is denoted by d_ν, which describes the model complexity. Let X_ν be the $n \times d_\nu$ matrix whose columns correspond to the selected covariates in model ν, let β_ν be the d_ν-dimensional vector of parameters, and let $\mathcal{B}_\nu$ be the corresponding parameter space.

The Wilcoxon-type generalized BIC chooses the model that yields the smallest value of

$$W_n(\hat{\beta}_\nu) + \frac{\tau}{2} d_\nu \log n \tag{4}$$

where $\hat{\beta}_\nu$ minimizes the Wilcoxon-type dispersion function (3) with X_i' replaced by $X_{\nu i}'$, the ith row of X_ν, and $\tau = \{\sqrt{12} \int f^2(u)du\}^{-1}$ is a constant related to Wilcoxon analysis.

Recent developments in software and algorithms for Wilcoxon analysis of regression models make the implementation of the Wilcoxon-type generalized BIC as convenient as that of Schwartz's BIC. The key quantities underlying the computation are $\hat{\beta}_\nu$ and an estimator of τ, which can be obtained by the functions wwest and wilcoxontau in the R software developed by Terpstra & McKean (2005). Alternatively, $\hat{\beta}_\nu$ can be calculated by applying an iterated reweighted least squares algorithm of Sievers & Abebe (2004), which can be easily carried out with major software packages.

3·2. *A heuristic Bayesian derivation*

In the Bayesian framework, model comparison is based on posterior probabilities. Consider a set of candidate models $M_1, \ldots, M_m$. Assume that model M_i, $i = 1, \ldots, m$, has a prior probability $\pi(M_i)$, and that its parameter β_i has a prior distribution $\pi(\beta_i|M_i)$. Then the posterior probability of model M_i given data D satisfies

$$\begin{aligned} p(M_i|D) &\propto \pi(M_i)p(D|M_i) \\ &\propto \pi(M_i)\int p(D|\beta_i, M_i)\pi(\beta_i|M_i)d\beta_i. \end{aligned}$$

In practice, the candidate models are often assumed to be equally likely, so that $\pi(M_i)$ is taken to be constant. As a result, model assessment crucially depends on the integral $\int p(D|\beta_i, M_i)\pi(\beta_i|M_i)d\beta_i$, which is often called the integrated or marginal likelihood for model M_i. For any two candidate models, the ratio of their corresponding integrated likelihoods gives the Bayes factor, a number that evaluates the evidence in favour of one model over the other.

Schwartz's BIC is an approximation to the logarithm of the integrated likelihood, and there is a similar heuristic derivation for the Wilcoxon-type generalized BIC. We consider an artificial likelihood

$$L(D|\beta_i, M_i) \propto \exp\{-\mathrm{W_n}(\beta_\mathrm{i})/\tau\}, \tag{5}$$

where $W_n(\beta_i)$ is the Wilcoxon dispersion function (3) corresponding to model M_i. The main motivation for using (5) as an artificial likelihood is that $-W_n(\beta_i)/\tau$ shares some essential properties of a parametric loglikelihood. To see this more clearly, note that under the null hypothesis $\beta_i = \beta_{i0}$, the test statistic $2\{W_n(\beta_{i0}) - W_n(\hat{\beta}_i)\}/\tau$ asymptotically has a χ^2 distribution (McKean & Hettmansperger, 1976), as for the likelihood ratio test. Moreover, minimizing $W_n(\beta_i)$ gives the Wilcoxon estimator, which works just like maximizing a loglikelihood function. The artificial likelihood relaxes the parametric assumption of Schwartz's BIC by dropping the need to specify a parametric likelihood function. In similar spirit, Pettitt (1982) uses an approximation to the marginal likelihood of ranks in Bayesian inference of linear models, Lazar (2003) applied empirical likelihood in Bayesian

analysis and Zhan & Hettmansperger (2007) proposed a rank-based pseudo-likelihood in a two-sample location problem for Bayesian estimation and testing.

Next, we drop the subscript i and consider the integrated likelihood corresponding to the artificial likelihood, $\int \exp\{-\mathrm{W_n}(\beta)/\tau\}\pi(\beta|M)\mathrm{d}\beta$. We will approximate this integral, which is generally intractable, using the Laplace method (Tierney & Kadane, 1986; Raftery, 1996). The basic idea of Laplace approximation is that, with large sample size, the integral is largely determined by the value of the integrand in a region close to $\tilde{\beta}$, the value of β that maximizes $g_n(\beta) = -W_n(\beta)/\tau + \log\{\pi(\beta|M)\}$.

For Schwartz's BIC, the Laplace approximation is performed by a second-order Taylor expansion of $g_n(\beta)$ around $\tilde{\beta}$, but the same approach is not directly feasible for the Wilcoxon-type BIC because $W_n(\beta)$ is not differentiable everywhere. Most of the priors used in practice for β satisfy $\log\{\pi(\beta|M)\} = O(1)$, as in the case of the commonly used unit information prior (Kass & Wasserman, 1995; Raftery, 1996; Volinsky & Raftery, 2000). Thus, for a large sample, we have $\tilde{\beta} \doteq \hat{\beta}$, the Wilcoxon estimator. In a small neighbourhood around $\tilde{\beta}$, with probability approaching one the Wilcoxon dispersion function $W_n(\beta)$ can be uniformly approximated by the quadratic function

$$Q_n(\beta) = (2\tau)^{-1} n(\beta - \beta_0)'\Sigma(\beta - \beta_0) - (\beta - \beta_0)'S_n(\beta_0) + W_n(\beta_0),$$

where β_0 is the population parameter value of the model under consideration,

$$S_n(\beta) = \sqrt{12}\sum_{i=1}^{n}\left\{(n+1)^{-1}R(Y_i - X_i'\beta) - 1/2\right\}X_i$$

and $\Sigma = \lim_{n\to\infty} n^{-1}X'X$; see Hettmansperger & McKean (1998, §3·5).

Replace $g_n(\beta)$ by $h_n(\beta) = -Q_n(\beta)/\tau + \log\{\pi(\beta|M)\}$. Applying Laplace approximation, we obtain

$$\begin{aligned}\int \exp\{-\mathrm{W_n}(\beta)/\tau\}\pi(\beta|M)\mathrm{d}\beta &\doteq \exp\{\mathrm{h_n}(\tilde{\beta})\}\int \exp\left\{(\beta-\tilde{\beta})'\mathrm{h_n''}(\tilde{\beta})(\beta-\tilde{\beta})/2\right\}\mathrm{d}\beta \\ &= \exp\{\mathrm{h_n}(\tilde{\beta})\}(2\pi)^{\mathrm{d}/2}|\mathrm{A}|^{-1/2},\end{aligned}$$

where $A = -h_n''(\tilde{\beta})$ and d is the dimension of the model. The above approximation has a relative error of order $o(1)$ (Tierney & Kadane, 1986). We immediately obtain

$$\begin{aligned} & \log\left[\int \exp\{-\mathrm{W_n}(\beta)/\tau\}\pi(\beta|M)\mathrm{d}\beta\right] \\ = \; & -W_n(\tilde{\beta})/\tau + \log\{\pi(\tilde{\beta}|M)\} + (d/2)\log(2\pi) - (1/2)\log|A| + o(1). \end{aligned}$$

Simple calculation yields $h_n''(\tilde{\beta}) \doteq -n\tau^{-2}\Sigma$, and therefore $|A| \doteq n^d|\tau^{-2}\Sigma|$ and the second last term of the right-hand side above becomes $-(d/2)\log n - (1/2)\log(|\tau^{-2}\Sigma|)$. Note that $-W_n(\tilde{\beta})/\tau$ is of order $O(n)$, the next smaller term is $-(d/2)\log n$, all the other terms are of order $O(1)$ or less. If we ignoring terms of smaller order, finding the model that gives the highest posterior probability based on the artificial likelihood (5) reduces to minimizing the Wilcoxon-type generalized BIC defined in (4).

4. Asymptotic properties

4·1. *A general consistency property*

Schwarz's BIC is consistent in the sense that it selects the true model with probability approaching one if such a true model is in the class of candidate models, but only when the likelihood function is correctly specified (Nishii, 1984). The main result of this subsection establishes that the Wilcoxon-type generalized BIC is consistent without the need to impose any parametric distributional assumption.

Consider a class of finitely many candidate models, each indexed by a p-dimensional binary vector ν, as discussed in §3·1. Assume that this class contains the true model, which is indexed by ν_0. If a candidate model nests the true model, it is called a correct model. The collection of all correct models is denoted by $\mathcal{M}^c$. Then as the samples size increases the Wilcoxon-type generalized BIC identifies the most parsimonious model in $\mathcal{M}^c$, i.e., the

one indexed by ν_0, with probability approaching one. Thus under the general regularity conditions given in the Appendix, the Wilcoxon-type generalized BIC is consistent.

THEOREM 1. *Let the model selected by the Wilcoxon-type generalized* BIC *be indexed by* $\hat{\nu}$*, and assume that the conditions in the Appendix are satisfied. Then* $\mathrm{pr}(\hat{\nu} = \nu_0) \to 1$ *as* $n \to \infty$.

The proof of this theorem is given in the Appendix. The idea of the proof is similar to that of Nishii (1984). When comparing a correct model with an incorrect model, the first term of the criterion function (4), which measures the goodness-of-fit of the model, asymptotically dominates and the correct model is preferred; when comparing a simpler correct model with a more complex correct model, the second term of the criterion function, i.e, the penalty term, asymptotically dominates and the simpler model is preferred. Hence with probability approaching one, the Wilcoxon-type generalized BIC favours the true model over either an incorrect model or a correct but more complex model.

Without going into the debate on the fundamental philosophical issues such as the existence of a true model and the ultimate goal of statistical modelling, one may still want to know how the proposed procedure works when the true model is not in the class of candidate models. Burnham & Anderson (2002, §6.4.2) pointed out that if this happens then Schwarz's BIC is consistent for the so-called quasi-true model, which is the best model within the class of candidate models according to a measure based on the Kullback-Leibler distance. A similar conclusion also holds for the Wilcoxon-type BIC. A careful examination of the proof in the Appendix reveals that, when two incorrect models are compared, the first term of (4) asymptotically dominates. Thus, with probability approaching one the Wilcoxon-type BIC picks the quasi-true model, which is defined as in Burnham & Anderson except that we need to use a generalized notion of the Kullback-Leibler distance (Shi & Tsai, 1998), which defines the Kullback-Leibler distance between the candidate model ν and the true model as $E_0|U_i - U_j|$ with $U_i = Y_i - X'_{\nu i}\beta_\nu$ and E_0 being the expectation

evaluated under the true model.

4·2. *Comparison of the L_2 risk*

To further assess the relative goodness of the Wilcoxon-type generalized BIC and Schwarz's BIC, we next compare their performance in terms of their respective L_2 risk function. Our analysis below suggests that the Wilcoxon-type generalized BIC is an attractive alternative to Schwarz's BIC, especially when the distribution is heavy-tailed or contaminated by outliers.

Write the regression model in §2 as $Y = \mu + \epsilon$, where $\mu = \alpha 1_n + X\beta$. For a given model selection criterion δ, assume that it selects a candidate model indexed by $\hat{\nu}$. Then the corresponding L_2 risk is defined as

$$R_\delta\{\mu, \mu(\hat{\nu})\} = n^{-1}E[\{\mu - \mu(\hat{\nu})\}'\{\mu - \mu(\hat{\nu})\}],$$

where $\mu(\hat{\nu})$ is an estimator of μ based on the selected model. Procedures with smaller L_2 risk are to be preferred.

For ease of exposition, we consider an important special case in which ϵ has a density function symmetric about zero. Let $Z = (1, X)'$ and $\gamma = (\alpha, \beta')'$, and assume $n^{-1}Z'Z \to \Lambda$, a positive definite matrix, then the least squares estimator $\hat{\gamma}_{LS}$ satisfies $\sqrt{n}(\hat{\gamma}_{LS} - \gamma) \to N_{p+1}(0, \sigma^2\Lambda^{-1})$ in distribution, and the Wilcoxon estimator $\hat{\gamma}_W = (\hat{\alpha}_W, \hat{\beta}_W)'$ satisfies $\sqrt{n}(\hat{\gamma}_W - \gamma) \to N_{p+1}(0, \tau^2\Lambda^{-1})$ in distribution, with $\hat{\beta}_W$ minimizing $W_n(\beta)$ and $\hat{\alpha}_W = \text{median}\{(\hat{e}_i + \hat{e}_j)/2, i \le j\}$ being the median of the Walsh averages of the residuals $\hat{e}_i = Y_i - X\hat{\beta}_W$; see McKean & Hettmansperger (1978).

It is straightforward to check that the risk function of Schwarz's BIC converges in probability to $(p+1)\sigma^2$, and that of the Wilcoxon-type generalized BIC converges in probability to $(p+1)\tau^2$. We define the asymptotic relative efficiency of the Wilcoxon-type generalized BIC versus Schwarz's BIC as the ratio of their corresponding asymptotic L^2 risk functions,

i.e.,

$$\text{ARE}_{\text{W,LS}} = \sigma^2/\tau^2 = 12\sigma^2 \left\{ \int f^2(u)du \right\}^2 .$$

Larger values of $\text{ARE}_{\text{W,LS}}$ indicate higher efficiency. Not surprisingly, this asymptotic relative efficiency for model selection is the same as the well-known result for comparing the Wilcoxon test with the t-test for the one-sample location problem. The value of $\text{ARE}_{\text{W,LS}}$ depends on the underlying error distribution. Its value is generally significantly greater than one for heavier-tailed errors. For example, for the double-exponential distribution, $\text{ARE}_{\text{W,LS}}$ is 1.5; for the t_3 distribution, $\text{ARE}_{\text{W,LS}}$ is 1.9; and for the normal distribution, $\text{ARE}_{\text{W,LS}}$ is 0.955. Hence the superior performance of the Wilcoxon-type generalized BIC for heavier-tailed errors comes with little sacrifice of efficiency under normality.

Finally, we compare the performance of the Wilcoxon-type generalized BIC and Schwarz's BIC by considering a contaminated normal error distribution, with probability density

$$f(x) = (1-\delta)\phi(x) + \delta c^{-1}\phi(c^{-1}x),$$

where $\phi(x)$ is the standard normal distribution density function, $c > 1$, and $0 < \delta < 1$ determines the amount of contamination. The asymptotic relative efficiencies of the Wilcoxon-type generalized BIC versus Schwarz's BIC for $c = 3$ and $\delta = 0.00, 0.01, 0.03, 0.05, 0.10$ and 0.15 are respectively 0.955, 1.009, 1.108, 1.196, 1.373 and 1.497. The values are the same as those in Table 1.7.1 of Hettmansperger & McKean (1998) for comparing the Wilcoxon test with the t-test. When the amount of contamination is 10%, $\text{ARE}_{\text{W,LS}}$ is 1.373, and when the amount of contamination is 15%, $\text{ARE}_{\text{W,LS}}$ is 1.497.

5. Numerical simulations

To compare the practical effectiveness of the Wilcoxon-type generalized BIC and Schwarz's BIC, we report results from a Monte Carlo study.

The data are generated from

$$Y_i = \beta X_i + \sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where $X_i = (x_{i1}, \ldots, x_{i5})'$, $\beta = (1, 1, 0, 0, 1)'$, and the ϵ_i are independent and identically distributed random errors. For the covariates, $x_{i1} \equiv 1$, and $(x_{i2}, \ldots, x_{i5})'$ follows a multivariate normal distribution with mean zero and a covariance matrix that has marginal variances one and an AR(1) correlation structure with autocorrelation coefficient $\rho = 0.5$; that is $\text{corr}(x_{ij}, x_{ik}) = 0.5^{|j-k|}$, for $2 \leq j, k \leq 5$. We set $\sigma = 0.5$ and 1.5 to adjust the signal-to-noise ratio. Three different error distributions are considered: the standard normal distribution $N(0, 1)$; the t_4 distribution standardized to have mean zero and variance one; and the contaminated normal distribution with 90% of the data from $N(0, 1)$ and 10% from $N(0, 25)$.

Based on 500 simulation runs, we calculate the proportions of times the true model, an underfitted model, missing at least one covariate in the true model, and an overfitted model, containing all the covariates in the true model and at least one extra covariate not in the true model, are selected, respectively. The results are summarized in Table 1 for two different values of σ, the three different error distributions and sample size $n = 50$; the pattern of results for $n = 100$ is very similar and is not reported.

Put Table 1 about here

Table 1 suggests that, in terms of the probability of selecting the true model, the Wilcoxon-type generalized BIC performs as well as Schwarz's BIC when the random error has the normal distribution, it is slightly better than Schwarz's BIC when the error distribution is the t_4 distribution, and it performs significantly better than Schwarz's BIC when the data are contaminated with larger outliers. For the contaminated normal random errors,

when $\sigma = 1.5$, Schwarz's BIC selects the true model only about 44% of the time, whereas the Wilcoxon-type generalized BIC selects the true model about 77% of the time.

We next compare these two procedures by computing the mean squared error of the selected model in the above setting for two different sample sizes $n = 50$ and 100. The mean squared error is defined as

$$\text{MSE} = n^{-1} \sum_{i=1}^{n} (\mu_i - \hat{\beta} X_i)^2,$$

where μ_i is the ith true mean and $\hat{\beta}$ is the estimated parameter for the selected model. The value of MSE is computed and then averaged over 500 simulation runs. Table 2 reports the results.

Put Table 2 about here

Table 2 shows that MSE decreases as the sample size becomes larger or the signal-to-noise ratio becomes smaller. For normal distribution, Schwarz's BIC results in smaller values of MSE , about $80 - 90\%$ of those of the Wilcoxon-type generalized BIC. However, for the t_4 distribution and the contaminated normal distribution, the Wilcoxon-type generalized BIC can lead to a much smaller value of MSE.

6. Windmills data

To illustrate the application and further demonstrate the stability of the Wilcoxon-type generalized BIC, we consider the Windmills data from Weisberg (2005, §10.4.1.). We analyze the data collected in November, 2002 at a test site and four nearby long-term weather sites in Northern South Dakota. The response variable is CSpd, the calculated wind speed

in metres per second. The values of the response variable and 13 explanatory variables were recorded every six hours. There are 114 observations. Such data are important for determining the potential energy that can be produced by a wind farm.

Following Weisberg, we consider six candidate multiple regression models that correspond to different sets of covariates.

Model 1: this model contains the intercept and Spd_1, the wind speed at reference site 1 in metres per second.

Model 2: this model is the same as Model 1 but with a separate intercept and slope for each of the 16 bins determined by the wind direction at reference site 1.

Model 3: this model contains the intercept, Spd_1, $\cos(\text{Dir}_1)$, $\sin(\text{Dir}_1)$ and the interaction terms $\text{Spd1} * \cos(\text{Dir}_1)$ and $\text{Spd1} * \sin(\text{Dir}_1)$, where Dir1 is the wind direction θ at reference site 1 in degrees.

Model 4: this model contains the intercept, Spd_1 and Spd_1Lag_1, where Spd_1Lag_1 is the wind speed at reference site 1 six hours previously.

Model 5: this model contains the intercept, Spd_1, Spd_2, Spd_3 and Spd_4, where Spd_i is the wind speed at reference site i, $i = 1, \ldots, 4$.

Model 6: this model contains the intercept, Spd_1, Spd_2, Spd_3, Spd_4, Spd_1Lag_1, Spd_1Lag_2, Spd_1Lag_3 and Spd_1Lag_4, where Spd_1Lag_i is the wind speed at reference site i six hours previously, $i = 1, \ldots, 4$.

The results based on Schwarz's BIC and the Wilcoxon-type generalized BIC are summarized in the first two columns of Table 3. The two approaches give almost the same ordering of the six models. In particular, with both methods they choose the same three top models: Model 5 is ranked the best, Model 6 comes the second and Model 1 comes the third.

Put Table 3 about here

There is still good reason to prefer the Wilcoxon-type generalized BIC. In fact, Schwarz's BIC may be unstable under minor disturbance of the data. To see this, we artificially change the value of the 25th response variable from 6.3 to 30. This disturbance creates an outlier in the right-hand tail. However, this outlier is not extreme as it is within 3 times the inter-quartile range of the nearest endpoint of the data: such an outlier is often classified as a mild outlier. Under this minor disturbance, the Wilcoxon-type generalized BIC ranks the six candidate models exactly as before. In contrast, Schwarz's BIC significantly changes its preference; it now picks Model 1 as the best model, Model 5 as the second best and Model 2 next. The details are given in Table 3.

Finally, we provide the least-squares fit and the Wilcoxon fit of Model 5, the best model selected by both criteria, in Table 4. The two methods yield very similar results.

Put Table 4 about here

7. Discussions

We may also develop a Wilcoxon-type AIC by working with a generalized notion of the Kullback-Leibler information. Both AIC and BIC require a search over possibly a large number of subsets of covariates. An alternative approach is to associate the Wilcoxon-type dispersion function with the Lasso penalty (Tibshirani, 1996) or the smoothly clipped absolute deviation penalty (Fan & Li, 2001). This would lead to a new procedure that selects variables and estimates parameters simultaneously by automatically shrinking small coefficients to zero. This dramatically reduces the computational cost and has the potential

to deal with the case where the number of covariates goes to infinity (Fan & Peng, 2004). These are topics of ongoing research.

The derivation in §3 suggests that the relative importance of M_i among the set of candidate models can be evaluated by $\exp(-\frac{1}{2}\text{WBIC}_i/\tau)\{\sum_{j=1}^{m}\exp(-\frac{1}{2}\text{WBIC}_j/\tau)\}^{-1}$, $i = 1,\ldots,m$, where WBIC_i denotes the Wilcoxon-type generalized BIC criterion function for model M_i. Admittedly, the Bayesian argument in this paper is heuristic and the properties of the resulted model selection procedure are mainly investigated from the frequentist perspective. Whether or not the use of the artificial likelihood can lead to valid Bayesian inference such as posterior estimation and whether or not it relates to any fully Bayes semiparametric approach needs further study.

Acknowledgement

I would like to thank the editor, an associate editor, an anonymous referee, Edsel Pena and Vance Berger for their valuable and constructive comments. This research was supported by a grant from the U.S. National Science Foundation.

Appendix

Assumptions and proof of Theorem 1

We need the following assumptions for Theorem 1.

Assumption A1. For the candidate model indexed by ν, $\lim_{n\to\infty} n^{-1}X_\nu'X_\nu \to \Sigma_\nu$, a $d_\nu \times d_\nu$ positive definite matrix.

Assumption A2. Let $H_\nu = X_\nu(X_\nu'X_\nu)^{-1}X_\nu'$ and let H_ν^{nii} be the ith diagonal entry of

this $n \times n$ matrix. Then $\lim_{n\to\infty} \max_{1\le i\le n} H_\nu^{nii} = 0$.

Assumption A3. The class of candidate models contains the true model, which is indexed by ν_0 and the parameter $\beta_{\nu 0}^*$. For this true model, the random errors $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with absolutely continuous probability density function $f(x)$, which has finite Fisher information, i.e., $\int \{f(x)\}^{-1} f'(x)^2 dx < \infty$.

Assumption A4. Let $G_n(\beta_\nu) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n |U_i - U_j|$, where $U_i = Y_i - X_{\nu i}' \beta_\nu$, and assume that $G(\beta_\nu) = \lim_{n\to\infty} G_n(\beta_\nu)$ has a unique minimizer β_ν^*. If the candidate model indexed by ν is an incorrect model, then $\beta_\nu^* \neq \beta_{\nu 0}$ in the sense that they are not equal when both are augmented to the p-dimensional vector of coefficients by filling zeros in the positions corresponding to covariates not included in the candidate model.

The above assumptions are similar to those in Hettmansperger & McKean (1998), which guarantee the asymptotic normality of the Wilcoxon rank estimator. For a candidate model indexed by ν, minimizing the Wilcoxon dispersion function is equivalent to minimizing $G_n(\beta_\nu)$ since $W_n(\beta_\nu) = \sqrt{3} n G_n(\beta_\nu)/2 + O(n^{-1})$. For the true model, $G(\beta_{\nu_0}) = \lim_{n\to\infty} n^{-2} \sum_{i=1}^n \sum_{j=1}^n |\epsilon_i - \epsilon_j - (X_{\nu_0 i} - X_{\nu_0 j})'(\beta_{\nu_0} - \beta_{\nu 0}^*)|$ has a unique minimizer $\beta_{\nu_0} = \beta_{\nu 0}^*$ because $\epsilon_i - \epsilon_j$ has median zero. For an incorrect candidate model, when Assumption A4 is satisfied, as in White (1981), it can be shown that the Wilcoxon estimator is consistent for β_ν^* and asymptotically normal.

Proof of Theorem 1. The proof consists of two steps.

Step 1. Consider any two candidate models, one correct and the other incorrect. Assume that the correct model is indexed by ν_1 and the incorrect model is indexed by ν_2. We shall show that with probability approaching one the Wilcoxon-type generalized BIC favours the correct model indexed by ν_1.

For simplicity, we use WBIC(ν) to denote the Wilcoxon-type generalized BIC criterion

function for a candidate model indexed by ν. Consider

$$\frac{1}{n}\{\text{WBIC}(\nu_2) - \text{WBIC}(\nu_1)\} = \frac{1}{n}\{W_n(\hat{\beta}_{\nu_2}) - W_n(\hat{\beta}_{\nu_1})\} - \frac{d_{\nu_1} - d_{\nu_2}}{2}\frac{\log n}{n},$$

where $\hat{\beta}_{\nu_i}$ is the Wilcoxon estimator for the candidate model indexed by ν_i, $i = 1, 2$. The Wilcoxon dispersion function can be locally approximated by a quadratic function; that is, for any constant $c > 0$,

$$\text{pr}\left\{\sup_{||\beta_\nu - \beta^*_\nu|| \le c/\sqrt{n}} |W_n(\beta_\nu) - U_n(\beta_\nu)| \ge \epsilon\right\} \to 0, \tag{A1}$$

as $n \to \infty$, where

$$U_n(\beta_\nu) = \frac{n}{2}(\beta_\nu - \beta^*_\nu)' B_\nu (\beta_\nu - \beta^*_\nu) - (\beta_\nu - \beta^*_\nu)' S_n(\beta^*_\nu) + W_n(\beta^*_\nu), \tag{A2}$$

with $B_\nu = \sqrt{12}\text{lim}_{n\to\infty} n^{-2}\sum_{i=1}^n \sum_{j=1}^n \left[(X_{\nu i} - X_{\nu j})' X_{\nu i} \int f\{u + \Delta(X_{\nu i}) - \Delta(X_{\nu j})\} f(u) du\right]$ and $\Delta(X_{\nu i}) = X'_{\nu_0 i}\beta_{\nu_0} - X'_{\nu i}\beta_\nu$. The approximation (A1) generalizes (3.5.12) of Hettmansperger & McKean (1998) in the spirit of White (1981) to allow for possible model misspecification. For the true model, $B_{\nu_0} = \tau^{-1}\Sigma_{\nu_0}$ and (A2) reduces to (3.5.12) of Hettmansperger & McKean (1998). Since $\hat{\beta}_{\nu_i}$ is consistent for $\beta^*_{\nu_i}$, with probability approaching one,

$$\begin{aligned}
&\frac{1}{n}\{\text{WBIC}(\nu_2) - \text{WBIC}(\nu_1)\} \\
=&\left\{\frac{1}{2}(\hat{\beta}_{\nu_2} - \beta^*_{\nu_2})' B_{\nu_2}(\hat{\beta}_{\nu_2} - \beta^*_{\nu_2}) - \frac{1}{n}(\hat{\beta}_{\nu_2} - \beta^*_{\nu_2})' S_n(\beta^*_{\nu_2}) + \frac{1}{n} W_n(\beta^*_{\nu_2})\right\} \\
&- \left\{\frac{1}{2}(\hat{\beta}_{\nu_1} - \beta^*_{\nu_1})' B_{\nu_1}(\hat{\beta}_{\nu_1} - \beta^*_{\nu_1}) - \frac{1}{n}(\hat{\beta}_{\nu_1} - \beta^*_{\nu_1})' S_n(\beta^*_{\nu_1}) + \frac{1}{n} W_n(\beta^*_{\nu_1})\right\} \\
&- \frac{d_{\nu_1} - d_{\nu_2}}{2}\frac{\log n}{n}.
\end{aligned} \tag{A3}$$

In the above expression $(\hat{\beta}_{\nu_i} - \beta^*_{\nu_i})' B_{\nu_i}(\hat{\beta}_{\nu_i} - \beta^*_{\nu_i})$ and $n^{-1}(\hat{\beta}_{\nu_i} - \beta^*_{\nu_i})' S_n(\beta^*_{\nu_i})$ are both $O_p(n^{-1})$. Thus, asymptotically $n^{-1}\{\text{WBIC}(\nu_2) - \text{WBIC}(\nu_1)\}$ is dominated by $n^{-1}\{W_n(\beta^*_{\nu_2}) - W_n(\beta^*_{\nu_1})\}$, which converges to $\sqrt{3}\{G(\beta^*_{\nu_2}) - G(\beta^*_{\nu_1})\}/2 > 0$ in probability, since ν_1 indexes a correct model. Therefore $\text{pr}\{\text{WBIC}(\nu_2) - \text{WBIC}(\nu_1) > 0\} \to 1$; that is, the Wilcoxon-type generalized BIC prefers the correct model with probability approaching one.

Step 2. Now consider comparing a model indexed by ν_1 with another model indexed by ν_2, both are correct but the one indexed by ν_1 is simpler.

Similarly to Step 1, with probability approaching one, we have expression (A3) for $n^{-1}\{\text{WBIC}(\nu_2)-\text{WBIC}(\nu_1)\}$. Now $(\hat{\beta}_{\nu_i}-\beta^*_{\nu_i})'B_{\nu_i}(\hat{\beta}_{\nu_i}-\beta^*_{\nu_i})$ and $n^{-1}(\hat{\beta}_{\nu_i}-\beta^*_{\nu_i})'S_n(\beta^*_{\nu_i})$ are both $O_p(n^{-1})$. Furthermore, $W_n(\beta^*_{\nu_1}) = W_n(\beta^*_{\nu_2})$ thus cancel each other, because both models are correct, and $X'_{\nu_1 i}\beta^*_{\nu_1} = X'_{\nu_2 i}\beta^*_{\nu_2}$, for all i. Therefore, asymptotically the dominating term of $n^{-1}\{\text{WBIC}(\nu_2) - \text{WBIC}(\nu_1)\}$ is

$$-\frac{d_{\nu_1} - d_{\nu_2}}{2}\frac{\log n}{n} > 0.$$

Thus with probability approaching one the Wilcoxon-type generalized BIC favours the simpler correct model indexed by ν_1. □

References

Broman, K. W. & Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Statist. Soc.* B **64**, 641-56.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer-Verlag.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348-60.

Fan, J. & Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928-61.

Hettmansperger, T. P. & McKean, J. W. (1983). A geometric interpretation of inferences based on ranks in the linear model. *J. Am. Statist. Assoc.* **78**, 885-93.

Hettmansperger, T. P. & McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of

residuals. *Ann. Math. Statist.* **43**, 1449-58.

Jureăková, J. (1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42**, 1328-38.

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion. *J. Am. Statist. Assoc.* **90**, 928-34.

Konishi, S., Ando, T. & Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**, 27-43.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika* **90**, 319-26.

McKean, J. W. & Hettmansperger, T. P. (1976). Tests of hypotheses of the general linear models based on ranks. *Commun. Statist.* A **5**, 693-709.

McKean, J. W. & Hettmansperger, T. P. (1978). A robust analysis of the general linear model based on one-step R-estimates. *Biometrika* **65**, 571-9.

McKean, J. W. & Schrader, R. M. (1980). The geometry of robust procedures in linear models. *J. R. Statist. Soc.* B **42**, 366-71.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758-65.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *J. R. Statist. Soc.* B **44**, 234-43.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika* **83**, 251-66.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

Shi, P. & Tsai, C.-L. (1998). A note on the unification of the Akaike information criterion. *J. R. Statist. Soc.* B **60**, 551-8.

Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika* **91**, 785-800.

Sievers, G. L. & Abebe, A. (2004). Rank estimation of regression coefficients using iterated reweighted least squares. *J. Statist. Comp. Simul.* **74**, 821-31.

Terpstra, J. & McKean, J. (2005). Rank-Based analysis of linear models using R. *J. Statist. Soft.* **14**, issue 7.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc.* B **58**, 267-88.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82-6.

Volinsky, C. T. & Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256-62.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edition. Hoboken NJ: John Wiley.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Am. Statist. Assoc.* **76**, 419-33.

Zhan, X. & Hettmansperger, T.P. (2007). Bayesian R-estimates in two-sample location models. *Comp. Statist. Data Anal.* **51**, 5077-89.

Table 1: Simulation study. The proportions of times Schwarz's BIC and the Wilcoxon-type generalized BIC select a true model, an underfitted model and an overfitted model, respectively, out of 500 simulation runs for two different values of σ and three different error distributions.

		Schwarz's BIC			Wilcoxon BIC		
Error distribution	σ	true	underfitted	overfitted	true	underfitted	overfitted
$N(0,1)$	0.5	0.89	0.00	0.11	0.90	0.00	0.10
	1.5	0.85	0.05	0.10	0.86	0.06	0.08
t_4	0.5	0.90	0.00	0.10	0.92	0.00	0.08
	1.5	0.85	0.04	0.11	0.91	0.01	0.08
$0.9N(0,1)$	0.5	0.88	0.00	0.12	0.92	0.00	0.08
$+0.1N(0,25)$	1.5	0.44	0.50	0.06	0.77	0.14	0.09

Table 2: Simulation study. The mean squared error for the model selected by Schwarz's BIC and that selected by the Wilcoxon-type generalized BIC, for two different sample sizes, two different values of σ and three different error distributions.

		$n = 50$		$n = 100$	
Error distribution	σ	Schwarz's BIC	Wilcoxon BIC	Schwarz's BIC	Wilcoxon BIC
$N(0,1)$	0.5	0.017	0.020	0.009	0.010
	1.5	0.175	0.204	0.077	0.094
t_4	0.5	0.017	0.014	0.008	0.007
	1.5	0.184	0.120	0.071	0.058
$0.9N(0,1)$	0.5	0.066	0.027	0.032	0.014
$+0.1N(0,25)$	1.5	0.903	0.379	0.435	0.138

Table 3: Windmills data. Comparison of the six candidate models in terms of Schwarz's BIC and the Wilcoxon-type generalized BIC when the data are without/with disturbance.

	Original data		Data under disturbance	
	Schwarz's BIC	Wilcoxon BIC	Schwarz's BIC	Wilcoxon BIC
Model 1	167.8	224.2	243.6	255.3
Model 2	174.1	230.8	251.8	262.1
Model 3	178.4	238.5	258.4	269.3
Model 4	172.4	233.5	248.0	264.8
Model 5	143.3	204.3	247.9	240.5
Model 6	155.3	211.6	261.1	248.3

Table 4: Windmills data. Estimated coefficients and standard errors for Model 5 as obtained by the least-squares method and the Wilcoxon method.

	Least-squares method		Wilcoxon method	
Covariate	estimate	std. error	estimate	std. error
intercept	1.83	0.41	1.71	0.45
Spd_1	0.43	0.10	0.41	0.10
Spd_2	0.22	0.10	0.22	0.10
Spd_3	0.03	0.12	0.03	0.12
Spd_4	0.23	0.12	0.25	0.12